Multi-ego-centered communities

Maximilien Danisch, Jean-Loup Guillaume, Benedicte Le Grand

Multi-Egocentered Communities

Maximilien Danisch, Jean-Loup Guillaume and Bénédicte Le Grand

Abstract

The community structure of a graph is defined in various ways in the literature: (i) partition, where nodes can belong to only one community. This vision is unrealistic and may lead to poor results because most nodes belong to several communities in real-world networks; (ii) overlapping community structure, which is the most natural view, but is often very difficult to identify in practice due to the complex structure of real-world networks and the huge potential number of such communities; (iii) egocentered community structure which focuses on individual nodes' communities and seems to be a good compromise.

In this chapter, the third vision is investigated; a new proximity measure based on opinion dynamics is proposed to score and select nodes according to their proximity to a node of interest. We call it the *carryover opinion*. In addition to be parameter-free, the carryover opinion can be calculated in a very time-efficient way and can thus be used in very large graphs.

We also go further in the idea of egocentered communities by introducing the new concept of *multi-egocentered communities*, i.e., focusing on the communities of a set of nodes rather than of a single node. A key idea is that, although one node generally belongs to numerous communities, e.g., friends, colleagues, family, a small set of appropriate nodes can fully characterize a single community. We also show how to unfold all egocentered communities of a given node using this notion of multi-egocentered community.

1 Introduction

In social networks, communities are groups of users who share common features or have similar interests; studying the community structure has thus many applications for advertising as well as market research. Given a set of users, the most common way of identifying communities consists in classifying them in identified or unknown classes; this is what traditional classification and clustering approaches do, e.g., k-means or others.

In the context of graphs, community detection generally aims at finding a partition of nodes, which means that each node belongs to one and only one community. However, if we consider social networks, where edges may represent friendship between users, it is hard to conceive that a user belongs to only one group: he/she clearly belongs to numerous groups, e.g., his/her

family, colleagues, various groups of friends. In order to be consistent with this, overlapping communities should be allowed. However, computing all these overlapping groups in a network leads to numerous problems; in particular, the number of potential groups in a network is 2^n, where n is the number of nodes: in addition to the time and space complexity of the algorithm, the interpretation of obtained results may be very difficult.

An interesting compromise is to focus on the groups related to one node. This type of communities is referred to as *egocentered communities*. For this task, we suggest to adopt a novel approach based on proximity between nodes instead of a cost function approach, as commonly seen in the literature, which suffers from local minimum and hidden scale parameter.

Even though we obtain interesting results, in some cases, egocentered community detection is still a difficult problem because a single node can still belong to numerous groups (up to 2^{n-1}); we therefore suggest to take into account the context by identifying the communities of a set of nodes, called *multi-egocentered communities*. In particular, we show that a small set of nodes is generally sufficient to define a unique community, which is generally not the case with one single node.

In addition to results obtained on small synthetic networks and small real-world networks, we have worked on a very large network which is a wikipedia dataset containing more than 2 million labeled pages and 40 million links, **?**.

This chapter details four recent contributions to the state of the art:

1. A new proximity measure between nodes based on opinion dynamics, which we call *the carryover opinion*. This proximity measure is parameter-free, takes into account the whole graph and not only a local view and is very fast to compute: the algorithm is in $O(te)$, where e is the number of edges and t is relatively small. (Calculating the proximity between one given node and all other nodes takes only few seconds for the whole wikipedia dataset).

2. The possibility of characterizing a node in terms of its egocentered community structure, i.e., stating whether it is in the center of a community or more peripheral and between several communities, thanks to the carryover opinion and the time-efficiency of its computation.

3. The new concept of multi-egocentered communities: communities related to a set of nodes, which extends the already established concept of egocentered communities.

4. An algorithm that unfolds all egocentered communities of a given node through unfolding multi-egocentered communities on the node of interest and some other carefully selected nodes.

The first section being the introduction, the following of this chapter is organized in five sections: the second section is a state of the art of community detection algorithms and node proximity measures for community detection. The third

section presents a new proximity measure, called carryover opinion, and its applications for the detection of egocentered communities. The fourth section shows how to use the carryover opinion to unfold multi-egocentered communities with some validations on real graphs. The fifth section details the algorithm that unfold all egocentered communities of a given node. Finally, the last section concludes and presents the perspectives for future works.

2 State of the Art

2.1 Community Detection

It has been found that most complex networks exhibit a community structure, ?. However, the concept of *community* itself is not well-defined. A common fuzzy definition is: a group of nodes more connected to one-another than to the nodes of the other groups. The idea of a community is also related to information propagation: information propagates faster within a community than through different communities. In most practical cases, communities are simply the output of an algorithm, without a more accurate definition.

As detailed in the introduction, even though the most realistic way of seeing the community structure is to consider overlapping communities, most initiatives in community detection applicable to very large graphs (i.e., dozens thousand nodes) are limited to the identification of a partition of nodes. A common way to unfold the community structure seen as a partition consists in maximizing a quality function, a popular one being modularity, ?. Even though maximizing this quality function is NP-hard, a good local minimum can be found very efficiently using the Louvain method, ?. Other approaches also exist, such as ?, where a metric based on random walks maps nodes into points in a Euclidean space, and thus transforms the problem of community detection into the one of clustering; the infomap method, ?, using techniques from data compression; or ?, using opinion dynamics, which is similar to the approach we will follow for egocentered communities.

There however exist algorithms to cope with the problem of overlapping community structure. The most popular is the k-clique percolation, ?, where a community is seen as a set of cliques of size k where each clique overlaps, at least, another one by k-1 nodes, where k is a parameter controlling the size of the cliques. Another interesting approach consists in partitioning the links instead of the nodes, which results in an overlapping community structure on nodes, ?. This can be done by applying the techniques established for communities seen as partition to the line-graph of the considered graph, ?. Another technique uses the non-determinism of algorithms for community seen as partition to obtain overlapping communities, ?.

Another trend in the literature related to the community structure focuses on one node. In addition to being a good compromise between the realism of overlapping communities and the feasibility of communities seen as a partition, this third way seems to have emerged because real networks, such as the Internet,

Facebook or the Web are huge and dynamic; in this context, it is hard to know the complete structure of the network, while it is still possible to know the structure around the neighborhood of one node. In the literature algorithms dealing with this problem consists in designing and optimizing a fitness function. Most of the time it is a function of the number of internal and external edges, ?????. Another work bases this fitness function on triangles, ?: the function, called Cohesion, compares the triangles made of three nodes within a community to triangles with only two nodes in the community and thus pointing out.

However, in addition to suffering from local minimum problems, these functions often have a hidden scale parameter. For instance Cohesion, incorporating a density of triangles term, decreases in $O(s^3)$ (where s is the number of selected nodes) in sparse graphs and thus leads to very small communities. This cost function is actually used to find *egommunities*, i.e., communities related to a node taking into account only its neighbors. In that case, since complex networks are not locally sparse, the density of triangles decreases slower and the function is less biased in favor of small size egommunities.

Another interesting algorithm based on fitness function is the one detailed in ?. The algorithm starts with all nodes in the community and by removing the nodes in the community, it greedily maximizes the minimum degree of the subgraph induced by the nodes in the community. Even though the algorithm is greedy, it is proved to reach a global optimum, however while the other algorithms are biased towards small size communities this one favors very big communities.

Because of the local minimum problems and since an unbiased cost function (with regard to scale) remains very hard to define, we suggest to use a proximity-based approach. The principle of our method can be split into three consecutive steps:

1. Calculate the proximity between the node of interest and all other nodes.
2. Rank nodes in decreasing proximity order, with regard to the node of interest.
3. Find irregularities in the decrease, if they exist, that can be due to the community structure.

2.2 Node Proximity Measure

Even though using a proximity measure (or metric) on nodes approach is novel for the study of egocentered communities, proximity measures have already been used for community detection seen as partition. For instance ? developed a metric based on random walks to map nodes into points in a Euclidean space. They thus transformed the problem of community detection into the one of clustering. They then used an agglomerative clustering algorithm to obtain a partition of nodes.

For our problem, various existing proximity measures or metrics on nodes may be used. However they all have one of the three following drawbacks: (i)

they are too restrictive, or (ii) they need an a priori parameter, or (iii) they are too slow to be computed for huge graphs. A selection of commonly-used proximity measures or metrics is presented in the following:

- Number of hops between nodes. This metric is too restrictive since it takes integer values which are small with regard to the size of the graph. It falls in category (i).

- Probability for a random walker who started to walk from the picked node to be on a given node after t iterations, **?**. This metric depends on the parameter t and belongs to category (ii). Moreover it gives an advantage to high degree nodes.

- Jaccard similarity coefficient. For 2 nodes a and b it is given by

$$J(a,b) = \frac{|N_a \cap N_b|}{|N_a \cup N_b|}$$

  where N_a (resp. N_b) is the set of the neighbors of a (resp. b). However, two nodes that do not share any neighbor have a proximity equal to zero. This is too restrictive for our problem and falls in category (i).

- Personalized page-rank, **?**, which is given by the following fix-point algorithm:

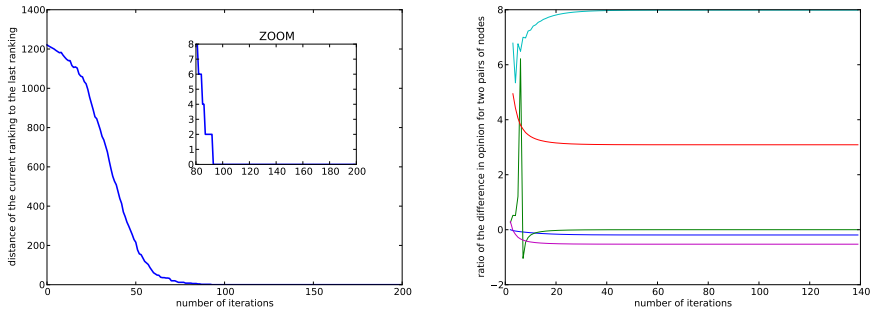
$$X_{t+1} = (1-\alpha)TX_t + \alpha X_0$$

  where X_t is the vector of the scores after n iterations, X_0 is initialized with the vector of all zeros except for the picked node which is set to one, T is the transition matrix: $T_{kl} = \frac{l_{kl}}{d_l}$, where l_{kl} is the weight of the link between nodes k and l, and d_l is the degree of node l. $\alpha \in]0,1[$ is a parameter which controls the depth of network exploration. The problem is that the result highly depends on α and gives an advantage to the nodes with a high degree. This proximity falls in category (ii).

- Hitting time (resp. commuting time) could be a solution. It is, for a source node and a target node, the expected number of steps that a random walker would take to go (resp. to go and come back) from the source to the target.

  With the node of interest as a target, all hitting times, i.e., with all nodes set alternatively as sources, can be calculated with a fix-point algorithm as detailed in **?**. However for very large graphs the fixed-point method is too slow to converge. Each iteration takes $O(e)$ (e, number of edges) and the number of iterations is about the maximum of the expected number of steps for all source nodes, which can be bigger than n (number of nodes). Thus this proximity falls in category (iii).

To our knowledge there is no proximity measure without at least one of the three identified drawbacks.



(a) (b)

Figure 1: Experiments validating conjectures 1 and 2. The experiments are carried out on the symmetrized polblogs network **?**, a network of 1222 nodes and 16717 edges. Figure **??** validates conjecture 1 by comparing the ranking of nodes according to their opinions to the ranking according to the last opinions obtained (for 200 iterations). As we can see, after only 95 iterations the ranking is not changing. The distance between the ranking we used is simply the number of misclassified nodes. Figure **??** validates conjecture 2 by plotting the ratio of the difference of two randomly chosen pairs of nodes. The experiment has been made 5 times, there are therefore five curves. As we can see, after only 40 iterations the ratio is quite constant, thus the differences in the opinion of a given pair of nodes is proportional to the one of any other pair.

3 A New Node Proximity Measure for Egocentered Communities

3.1 Carryover Opinion Metric

In this section, we define a proximity measure based on opinion dynamics, which takes into account the whole depth of the graph, is parameter free and is fast to compute.

Given a node of interest, the framework consists in first setting the opinion of this node to one and the opinion of all other nodes to zero. Then, at each time step, the opinion of every node is averaged with the one of this neighbors. The opinion of the node of interest is then reset to one. Its opinion thus does not change all along the process and remains equal to one (which means that the proximity between the node of interest and itself is one).

As such, this process is useless because it converges to an opinion of one for every node, however we have the feeling that nodes *closer* to the starting node will converge faster. The idea is to obtain a measure of that speed to characterize to what extent nodes are similar to the node of interest: the higher the speed the more similar the node. Two conjectures are needed to carry on:

Conjecture 1: after a sufficient number of iterations, the ranking of the nodes according to their opinion does not change anymore.

Conjecture 2: after a sufficient number of iterations, the difference between the opinion of two nodes decreases proportionally to the difference between the opinion of any other two nodes.[1]

The conjectures simply state that given four nodes a, b, c and d with opinion at iteration t noted O_a^t, O_b^t, O_c^t and O_d^t respectively. We have:

$$\lim_{t\to\infty} \frac{O_a^t - O_b^t}{O_c^t - O_d^t} = C_{a,b,c,d}$$

where $C_{a,b,c,d}$ is a constant depending only on nodes a, b, c and d.

These conjectures have been tested on various benchmarks and real-world networks with conclusive results. We show the results on figure **??**, where the experiment is carried out on the symmetrized polblogs network, **?**, a network of blogs and hyperlinks consisting in 1222 nodes and 16717 edges. As we can see, after a few iterations, the ranking of nodes according to their opinion does not change, while the difference between opinions becomes proportional.

It is thus possible to rescale the opinion at each iteration such that the lowest opinion is zero. The highest value is always one, which is the opinion of the node of interest. Scores between one and zero are thus obtained for each node at each iteration and the process converges towards a fix point. We call this value after convergence the carryover opinion, because even though the simple opinion process detailed above converges towards one for every nodes, the rescaling allows us to capture the proximity of nodes to the node of interest, which is carried over the whole process.

The node of interest beeing labeled i, each iteration thus consists in three steps:

$$\begin{array}{lcll} X_t & = & MX_{t-1} & \text{AVERAGING} \\ X_t & = & \frac{X_t - min(X_t)}{1 - min(X_t)} & \text{RESCALING} \\ X_t^i & = & 1 & \text{RESETTING} \end{array}$$

where:

- X_t is the score vector after t iterations and the component j of the vector X_t is noted X_t^j.
- X_0 is set the null vector, except for the node of interest, i, with value one.
- M is the averaging matrix, i.e., the transposed of the transition matrix : $M_{kl} = \frac{l_{kl}}{d_k}$, where l_{kl} is the weight of the link between the nodes k and l, and d_k is the degree of node k.

We tested the algorithm on the polblogs network, see figure **??**. After the convergence, which is nearly obtained after 40 iterations, the decrease in loglog

[1] Even though conjecture 2 implies conjecture 1, we think it is clearer to dissociate the two.

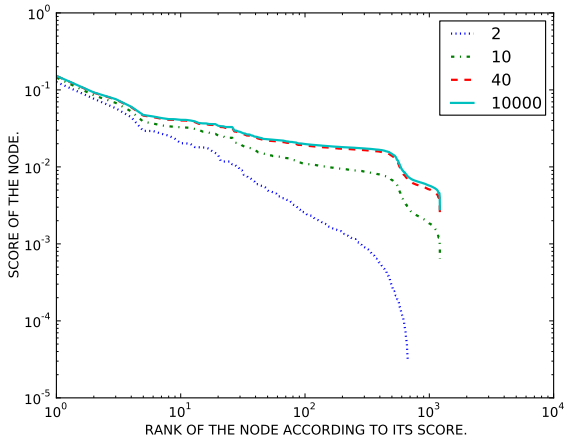


Figure 2: Experiment showing the convergence towards the carryover opinion. The experiment is carried out on the **?** polblogs network for which we randomly selected a node. The plot shows the score of each node as a function of its score ranking itself for 2, 10, 40 and 10000 iterations. Even though the order of nodes slightly changes during the first hundred iterations, as proved on figure **??**, the changes are negligible after 40 iterations.

scale is composed of two plateaus separated by a significant decrease in score values. This decrease appears around the 600th node. Actually the dataset contains 759 political blogs labeled as liberal and 443 labeled as conservative. In order to determine whether the nodes of the first plateau correspond to the picked node's community, we plotted the graph using the spring layout of **?**, using a circle (resp. square) shape for liberal (resp. conservative) blogs. The randomly picked node is located by an arrow. We then colored nodes according to their scores following a logarithmic scale, see in figure **??**. As we can see, the colors are consistent with labels: the randomly picked node was actually a liberal blog and most liberal blogs are dark while the conservative blogs remain white. When nodes are ranked in decreasing order according to the carryover opinion: 561 liberal nodes are among the 600 first ranked nodes, i.e., 93.5% of the 600 first ranked nodes are liberal; 617 liberal nodes are among the 759 first ranked nodes, i.e., $81,4\%$ of the 759 first ranked nodes are liberal.

We applied this technique to smaller networks, therefore easier to visualize. Interesting results have been obtained, as shown on figure **??**: Figure **??** shows the carryover opinion of nodes as a function of their carryover opinion ranking for a co-authorship network, **?**. The curve exhibits two major drops: the first one around the 50th node (the first 50 nodes therefore constitute the closest community of the picked node) and another one around the 180th (the first 180 nodes thus correspond to a larger community of the picked node, i.e., a community at a lower resolution). The corresponding nodes can be seen on the

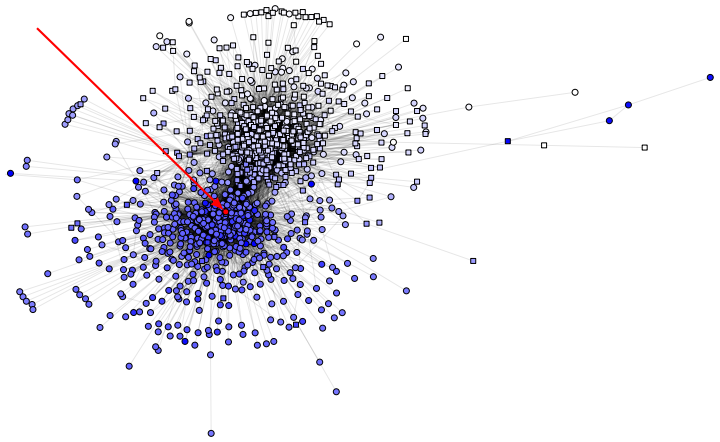
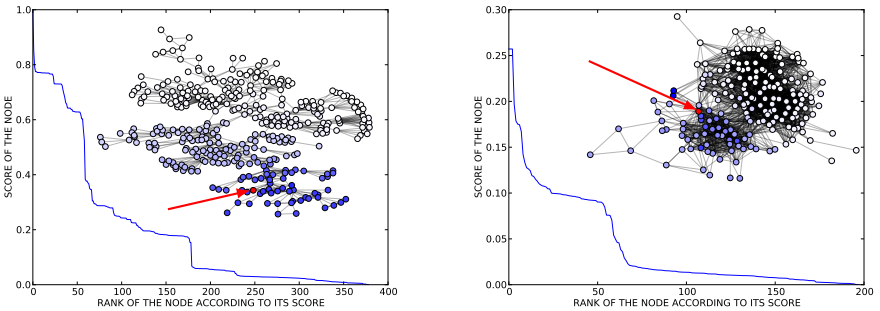


Figure 3: Drawing of the polblogs graph following the spring layout of **?**. The circles represent liberal blogs, while squares represent conservative blogs. An arrow points to the randomly picked node, while the higher the carryover opinion of a node, the darker its color, following a logarithmic scale.

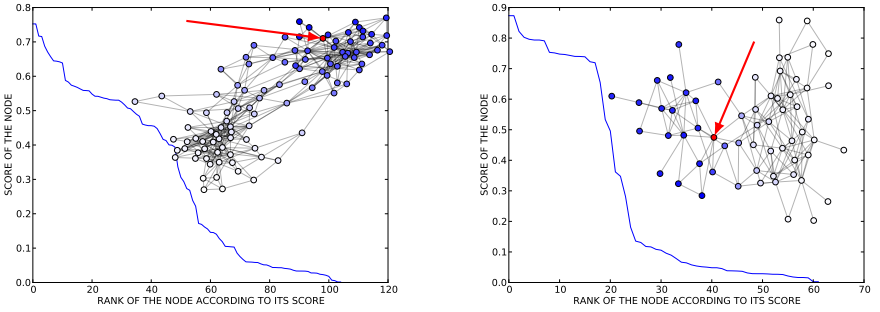
drawing where three different levels of color emerge. The succession of plateaus and decreases (on figures **??**, **??** and **??**) for three other networks also shows how useful the carryover opinion is to unfold egocentered communities.

As we can see on figure **??**, results obtained with the carryover opinion are not always the expected ones: this experiment has been carried out on a synthetic network consisting of three Erdos-Renyi graphs of hundred nodes with a link probability of 0.3, while nodes belonging to different Erdos-Renyi graphs have a probability of 0.05 to be linked. The value obtained for the first neighbors of the picked node somewhat dominates the community structure artificially generated, in fact the neighbors of the picked node have a high score even if they are in different Erdos-Renyi graphs. However one can argue that we are looking for the community(ies) of one node and, in that sense, if a node is linked to the picked node those two nodes already constitute a community. Actually the minimal value for a first neighbor with degree d is $\frac{1}{d}$, which makes sense: if all other neighbors of this first neighbor are *faraway* from the picked node, then this first neighbor is still $\frac{1}{d}$ part of the community(ies) of the picked node.

This effect (due to the communities of two nodes) can however be easily eliminated, as shown on figure **??**, by adding an additional step after the convergence of the carryover opinion: the picked node is removed from the graph and the value for each node is set to the average value of its neighbors. This



(a) (b)



(c) (d)

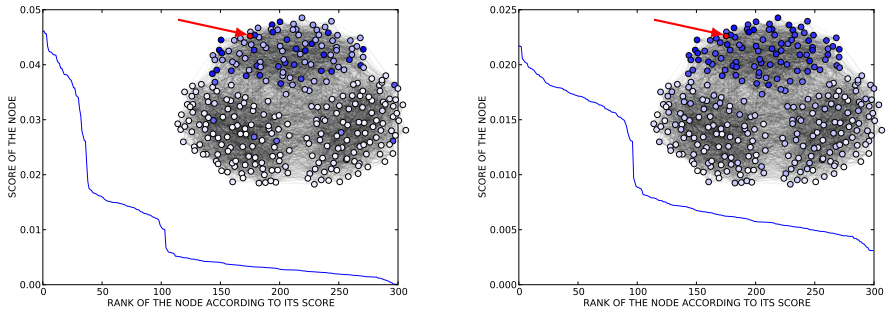
Figure 4: Result for four small visualizable networks. On the drawing of the networks, arrows point to the selected nodes, while the higher the score, the darker the node. The graphs are plotted using the graphviz layout. On small graphs a simple linear scale for the plot of the carryover opinion can be used. **??** is for a co-authorship network of 379 nodes and 914 edges, **?**. **??** is for a co-appearance network of jazz musicians of 198 nodes and 5484 edges, **?**. **??** is for a citation network of political books of 105 nodes and 441 vertices, **?**. **??** is for a social network of dolphins of 62 nodes and 159 edges, **?**.

affects only the first neighbors and it is the same as applying the transformation:

$$S = (S - \frac{1}{d})\frac{d}{d-1},$$

where S is the carryover opinion of a first neighbor.

We also can see that there are two effects that result in the final value of the carryover opinion: (i) 'a distance effect' and (ii) 'a redundancy effect' due to the community structure. As shown in figure **??**, the distance effect is sometimes dominating the redundancy effect. We argued that this is because the carryover opinion sees a pair of linked nodes as a community. The question is to know



(a) (b)

Figure 5: ?? shows the result for 3 Erdos-Renyi graphs (100,0.3), while nodes in different erdos-Renyi graphs are linked with probability 0.05. Figure ??, shows the same result, but with an additional step: the picked node is removed and the value for each node is set to the average value of its neighbors, i.e., a final averaging step is performed without the picked node. The higher the score, the darker the node.

how (if) this will affect the result for the nodes at distance two or more. To investigate this, we compare the decrease of the carryover opinion as a function of the distance for the wikipedia network (choosing the page 'boxing') and an Erdos-Renyi graph of the same average degree. As shown in figure ??, while on the Erdos-Renyi graph the decrease is exponential, on the wikipedia network only the neighbors of the picked node are affected. This means that there is no correlation between the distance and the value of the carryover opinion for nodes at distance two or more from the picked node. Thus this effect is only due to the fact that two linked nodes are considered as a community and the correcting step we suggested eliminates this effect.

Such an ideal structure of plateaus and strong decreases (as seen on figures ?? and figures ??) does not always appear. In fact it depends on two things: (i) the position of the picked node, i.e., central in a community or peripheral and thus within several communities. As shown on Figure ??, when the node is central the plateaus are clear while when the node is peripheral, no plateau is emerging. (ii) the structure of the community itself, i.e., the fact that community is well defined or not, as we can see on figure ??.

3.2 Egocentered Communities: Results on Large Graphs

The technique presented above does not need any a priori input parameter other than the graph and is very time-efficient. It can thus be used in huge graphs to find 'the community' or 'the communities' of a node if there is one, looking for various rates in the decrease. However, as already discussed, a node

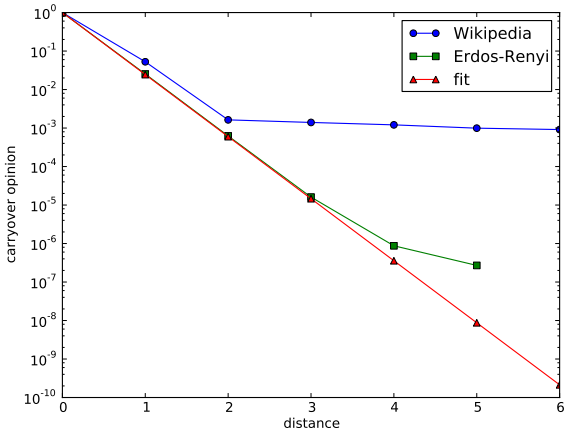
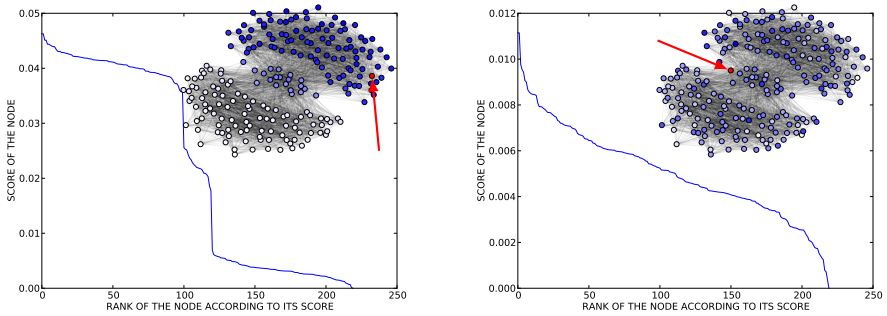


Figure 6: These plots show the average carryover opinion for nodes at a given distance from the node of interest as a function of the distance: Wikipedia is for the wikipedia network containing $n = 2,070,367$ nodes and $e = 42,336,614$ edges. Erdos-Renyi is for an Erdos-Renyi graph containing this same number of edges and nodes. Fit represents the curve $\frac{1}{degree^{distance}}$ where the degree is set to the average degree of the previous graph, i.e., $degree = \frac{2e}{n} = 40$



(a) (b)

Figure 7: Results given by the carryover opinion with the correcting step for two overlapping Erdos-Renyi graphs of 110 nodes with an edge probability of 0.3 overlapping on 20 nodes. The higher the score, the darker the node. As we can see on figure **??** when the picked node is at the center of a community the plateaus-decreases structure is clear, while it can be unclear when the node is peripheral, figure **??**.

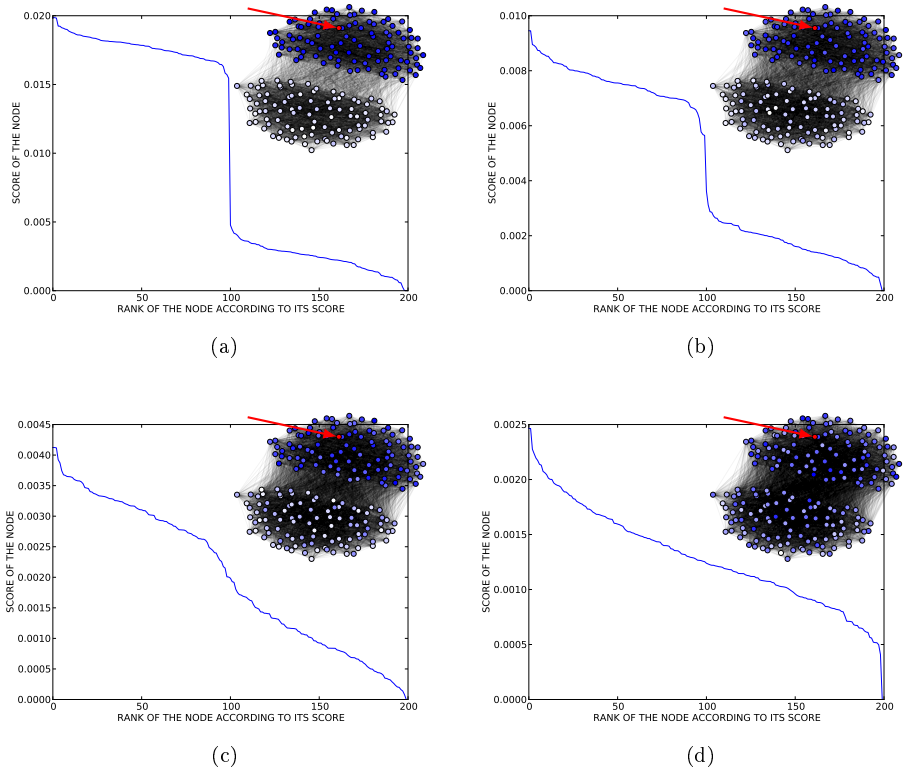


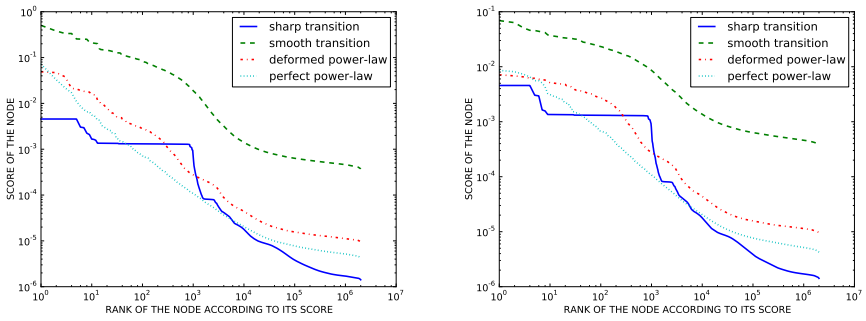
Figure 8: Results given by the carryover opinion with the correcting step for two Erdos-Renyi graphs (100,0.5). In Figure ?? (resp. ??, ??, ??) two nodes in different Erdos-Renyi graphs are linked with probability 0.1 (resp. 0.2, 0.3, 0.4).

often belongs to numerous communities and such a succession of plateaus and decreases is only occasionally observed.

Given randomly chosen nodes from the wikipedia network, figure ?? (resp. ??) shows the plots of the carryover opinion (resp. with the additional correcting step) for all nodes as a function of their ranking. The four types of curves show the four major trends one can obtain: sharp transition, smooth transition, deformed power law, perfect power law.

These four very different types of curves reflect very different structural properties of the nodes. Let us first notice that the correcting step does not modify much the curves, the bias due to communities of two nodes is thus minimal here. This may actually mean that there is only a little amount of weak ties (i.e., links between very different communities) in the wikipedia network.

Let us explain these four behaviors through analyzing the curves and the



(a) (b)

Figure 9: Plots of the carryover opinion of all nodes as a function of their ranking for four randomly picked nodes in the wikipedia network (left), and the same plots but after adding the correcting step (right). Sharp transition corresponds to the 'Cotton Township, Switzerland County, Indiana' node. Smooth transition corresponds to the 'Mafia' node. Deformed power law corresponds to the 'Mi-Hyun Kim' node. Perfect power law corresponds to the 'JNCO' node.

ranking of pages without the correcting step:

- The 'sharp transition' curve corresponds to the 'Cotton Township, Switzerland County, Indiana' page. As we can see the first 6 nodes constitute a plateau. These nodes correspond to the page 'Switzerland County, Indiana' and the 5 other townships of The Switzerland County. Then we withstand a decrease on the next 7 nodes which are tightly related to 'Township, Switzerland County' and 'Indiana'. The next 970 nodes constituting the second plateau all correspond to other townships in Indiana (with no exception, Indiana counting 1005 townships). The next decrease on about 1000 nodes is composed by nodes related to townships and Indiana and also a little about Illinois, while the following plateau on about 1000 additional nodes is composed of the pages of the townships of Illinois (with a few exceptions). The wavy decrease towards the final plateau smoothly transits towards far away related contexts, passing through Indiana related topics, Ohio's townships, Michigan's townships, other states townships, US related topics...

- The 'Smooth transition' curve is obtained for the 'Mafia' page. This node can characterize a community by itself: the first thousands pages are mafiosi names or organized crime related topics. However the community is more fuzzily defined than the ones for 'Cotton Township, Switzerland County, Indiana'.

- The 'Deformed power-law' curve is for 'Mi-Hyun Kim' page. The page is mainly linked to pages about Golf and Korea topics. The first thousand

pages are related to one or two of these topics, we obtain a superposition of the score of these topics, which leads to this wavy power law; this behavior is even clearer after applying the correcting step: we can then see two waves corresponding to a mixture of the two topics/communities (Korea and Golf).

- The 'perfect power-law' curve is for the 'JNCO' page, which is a clothing brand. As we can see the plot is a perfect power law that finishes with a low plateau. No community structure emerges from this plot; this is because the page is indeed linked to many different nodes that are part of various communities of different sizes fuzzily overlapping: 'JNCO' is linked to the pages 'Los Angeles', 'Jeans', 'Hip-hop', 'J.C. Penney', 'Graffiti', 'Kangaroo', 'Boxing', 'Nu Metal', from which hardly any context can emerge.

Concerning communities, we found that, in the same network, there seem to be two types of communities and we may characterize them as:

1. Well-defined communities, like the one of Switzerland country or Indiana.
2. Fuzzily defined communities, like the one of mafia.

Moreover, these communities can be multiscale: Switzerland country is a subcommunity of Indiana.
Concerning nodes, we found that, in the same network there are mainly three types of nodes (regarding communities):

1. Nodes which can, by themselves, define a community like 'Cotton Township, Switzerland County, Indiana' or 'mafia'.
2. Nodes which are in the middle of very few communities, like 'Mi-Hyun Kim'.
3. Nodes which are in a middle of a large number of communities, like 'JNCO'.

For a given node, these features can all be deduced from the shape of the curve representing their carryover opinion as a function of the ranking.

4 A New Vision of Communities

4.1 Multi-Egocentered Communities

It appears that, on the wikipedia network, most nodes have a -carryover opinion VS ranking- curve whose behavior is between deformed power-law and perfect power-law. Thus, in this network, nodes seem to belong to many communities; however, we have the intuition that a well chosen small set of nodes could define a single community.

The question is: how may the communities shared by a set of nodes be unfolded? We suggest to use the previously established proximity. The idea is that a node belonging to both, a community of $node_1$ AND a community of $node_2$ has to be similar to $node_1$ AND to $node_2$. The following example in figure **??** shows how to proceed:

1. Evaluate for all nodes the proximity to node1 and to node2.
2. The proximity to the set {$node_1$,$node_2$} is then given by the minimum, or by the geometric mean of the similarities to $node_1$ and the similarities to $node_2$. This quantity measures to what extent a node is near from $node_1$ AND $node_2$.[2]

The method is easily generalizable to a set of more than two nodes.

To validate the technique presented here, we extensively tested it and obtained good results on various homemade visualizable networks and on the Lancichinetti and Fortunato's benchmark for overlapping communities **?**. We present here the results for a particular trial on the benchmark: we built a network of 100,000 nodes with 10,000 nodes belonging to 3 communities and the others belonging to only one community, we used a mixing parameter of 0.2 and kept default values of power law coefficients for the degrees distribution and sizes of communities distribution. We picked two nodes belonging to three communities, one of each common to both of them. The results are presented on figure **??**: as we can see the unions of the three communities for both nodes is identified almost perfectly as is the community shared by both nodes. Indeed the Jaccard coefficient between the real communities and the one unfolded by the framework is always greater than 0.9.

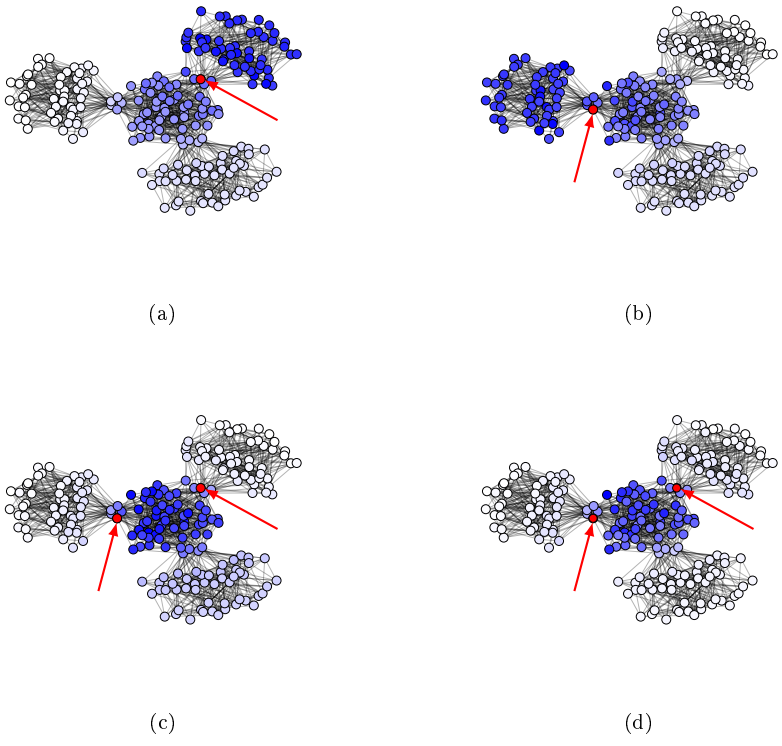
4.2 Multi-Egocentered Communities: Results on Large Graphs

We applied the framework described above to the wikipedia network using the minimum proximity of the picked nodes. Figure **??** shows the results for two nodes : 'Folk wrestling' and 'Torii school'. One is dedicated to the various types of traditional wrestling around the world, while the other one is dedicated to a traditional Japanese art school. Both curves are slightly deformed power-laws and do not uncover any community.

Figure **??** shows the result for 'Sumo' along with the minimum of the scores for the pages 'Folk wrestling' and 'Torii school' and the same rescaled minimum, such that it starts at 1.

As we can see the two curves have exactly the same structure: a plateau followed by a decrease at about the 350th node. 'Folk wrestling' and 'Torii school' are related to 'Sumo' in a transversal way. Doing the minimum of the scores for these two pages gives us a score of how nodes are related to 'Folk wrestling'

[2] Doing the arithmetic mean of the proximity or their maximum is not relevant for our problem, since this would unfold nodes that are part of a community of node1 OR node2.



(a) (b)

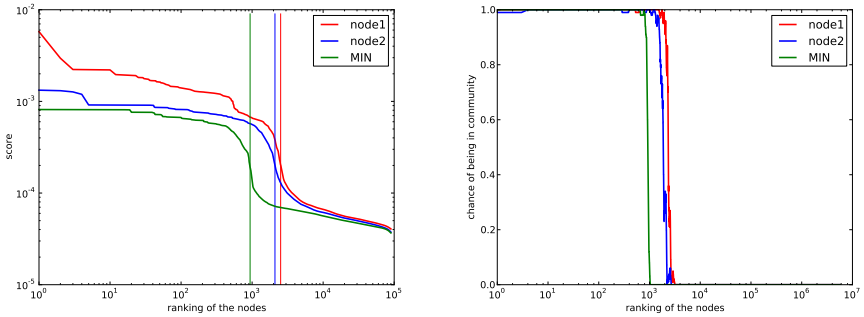
(c) (d)

Figure 10: Result for 4 overlapping Erdos-Renyi graphs of 50 nodes and an edge probability of 0.2 overlapping on 5 nodes. The darker a node, the higher its score. Arrows point to selected nodes. Figure **??** (resp. figure **??**) gives the (rescaled) minimum (resp. geometric mean) of the scores on the experiments presented on figures **??** and **??**. The community shared by both red nodes is emerging.

and Torii school' which actually correspond to 'Sumo'. Comparing the 350 first nodes of each experiments gives that:

- 14 nodes are in the first 350 nodes of 'Sumo' and 'Torii school',
- 12 nodes are in the first 350 nodes of 'Sumo' and 'Folk wrestling',
- 337 nodes are in the first 350 nodes of 'Sumo' and the minimum of 'Folk wrestling' and 'Torii school'.

Also, the node having the highest score when doing the minimum of the carry-over opinion for 'Folk wrestling' and 'Torii school' is actually 'Sumo'. In that case we found a set of pages which define a community already defined by a



(a) (b)

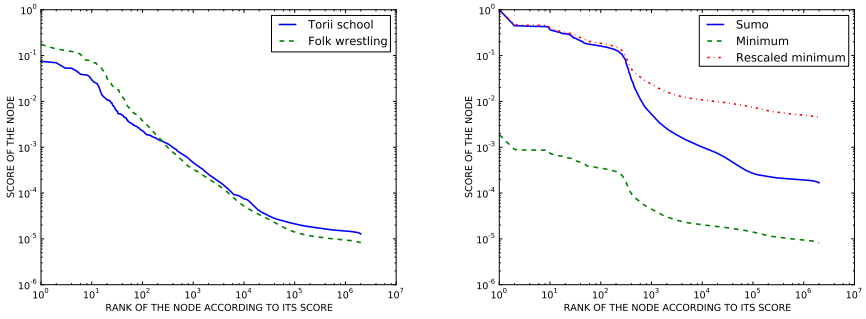
Figure 11: Figure ?? shows the carryover opinion of all nodes as a function of their ranking for the two nodes having three communities while sharing one (node 1 and node 2). It also shows the minimum of these two scores for all nodes as a function of the ranking (MIN). The highest slope of each curve is identified by a vertical bar. Figure ?? shows the proportion of nodes (on a sliding window containing 100 nodes) actually in one of the three communities, as well as the proportion of nodes actually in the shared community, as a function of the same rankings. We can see that the highest slopes correspond to the transition: "in the community/out of the community".

single node (the egocentered community of 'Sumo'), but we believe that it is also possible to find multi-egocentered communities which are not egocentered.

It seems that using the minimum of both values could be more effective, however doing the geometric mean can allow to weight the set (possibly weighting some nodes negatively) to better investigate the overlapping. Also, using the minimum may be less stable in large graphs, since a single node added to the initial set could highly change the result (for instance if a node that has nothing to do with the rest of the set is added). Conversely, adding a node very similar to a node already present in the set would not change the result. However, in our experiments, we obtained better results with the minimum than with the geometric mean.

5 How to Find All Egocentered Communities of a Given Node

In this section we propose an approach to find all egocentered communities of a given node through finding multi-egocentered communities of the node of interest and some other candidates. We show the result of our method when applied to a real large graph: the whole wikipedia network containing more than 2 million labeled pages and 40 million edges hyperlinks ?.



(a) (b)

Figure 12: Figure **??** shows the results for two nodes, 'Folk wrestling' and 'Torii school': two power laws. Figure **??** shows the result for 'Sumo' along with the minimum of the scores for the pages 'Folk wrestling' and 'Torii school' and the same rescaled minimum, such that it starts at 1.

5.1 Framework

Given a specific node u, we measure the proximity[3] of all nodes in the graph to u and then try to find irregularities in the decrease of these proximity values, as explained in the previous sections. Such irregularities can reflect the presence of one or more communities. However this routine often leads to a power-law with no plateau and from which no scale can be extracted; this happens when lots of communities of various sizes are overlapping which is often the case. To cope with this problem, we use the notion of multi-egocentered community (in particular bi-egocentered community), i.e., centered on a set of nodes instead of a single node. We thus need to smartly pick another node, v, evaluate the proximity of all nodes in the graph to v and then for each nodes in the graph, compute the minimum of the score obtained from u and the score obtained from v: this minimum evaluates how a node is similar to u AND v. Note that doing this sometimes leads to the identification of a community which does not contain u and/or v, however since we are interested only in communities containing u, we use v as an artifact and keep a community only if it contains u, regardless of v. The framework consists in doing this for enough candidate nodes v in order to obtain all communities of u. We will now detail the steps of the framework.

5.2 How to Chose the Candidates for v?

First, the carryover opinion of u has to be computed. This gives a real value for each node present in the graph: its proximity to u. Sorting the obtained values and plotting them as a function of their ranking leads to the carryover curve. If

[3]Even though other proximity measures can be used, we use the carryover opinion.

the outcome is a power-law, there is no relevant scale and u certainly belongs to several communities of various sizes.

We then want to pick v such that v and u roughly share exactly one community. If v is very dissimilar from u then it is very unlikely that u and v will share a common community: computing the minimum of the scores obtained from running the carryover opinion from u and the scores obtained from running the carryover opinion from v will lead to very small values. Indeed if the two nodes share no community, at least one of the scores will be very low. Conversely if v is extremely similar to u then the two nodes will share many communities. The carryover opinion values obtained from u and v will be roughly the same and doing the minimum will not give more information.

Thus v must be similar enough to u, but not too similar: it has to have a carryover score obtained from u not too high and not too low. A low and high proximity thresholds can be manually tuned to select all nodes at the right distance in order to fasten the execution.

It is quite likely that many of these nodes at the right distance will lead to the identification of the same community, therefore not all of them need to be candidates; a random selection of them can be used if the running time of the algorithm matters. More precise selection strategies will be discussed in the future work section.

5.3 How to Identify the Egocentered Community of u and v?

In order to identify the potential community centered on both u and v, we must compute the minimum of the carryover values obtained from u and from v for each node, w, of the graph. The minimum of the two scores is therefore a measure of the belonging of w to the community of both u and v. We can then sort these minimum values and plot the minimum carryover curve. As before, an irregularity in the decrease, i.e., a plateau followed by a strong decrease, indicates that all nodes before the decrease constitute a community.

Detecting a plateau followed by a strong decrease can be done automatically: if the maximum slope is higher than a given threshold, the nodes before this maximum slope constitute a community. This threshold should be manually tuned. If there are several sharp decreases, we currently only detect the sharpest. This could be improved in the future.

In addition, if u is before the decrease then u is in the community. In that case, these nodes before the decrease constitute a community of u. Note that v does not need to belong to this community since we are trying to find communities around u and that v is only a node that we use to find such communities.

As such this method is not very efficient when the carryover opinion is applied to a very high degree node connected to a very large number of communities. In that case, the carryover tends to give high values to every node in the graph and calculating the minimum with the scores obtained from a less popular node, which gives lower values to the nodes, will simply result in the values obtained

with this second node. A rescaling before doing the minimum can fix the problem. Indeed the lowest values obtained by running the carryover opinion result in a plateau, rescaling (in logarithmic scale) the values such that these plateaus are at the same level solves this problem.

5.4 Cleaning the Output and Labeling the Communities

The output of the two previous steps is a set of communities (where each node is scored), since each candidate node can yield a community. These communities need to be postprocessed, since many of them are very similar.

We propose to clean the output as follows: if the Jaccard similarity [4] between two communities (or any other similarity measure between sets) is too large, it means that although communities are actually the same, they appear to be different because of the noise. In that case we only keep the intersection of these two communities. For each node in this new (intersection) community, the score is given by the sum of the scores in the original communities.

We perform an optional cleaning step, which enhances the results: if a community is dissimilar to all other communities, we simply remove it. Indeed, a good community should appear for different candidate nodes. We observed that such communities come from the detection of a plateau/decrease structure which does not exist (it often happens when the threshold is not set to a proper value).

We finally label the community with the label of the best ranked node in the community, i.e., the node whose sum of values is the highest. If two communities have the same label we suggest to keep both (it can be different scales of the same community).

This algorithm finally returns a set of distinct labeled communities. We now show some results on a real network.

5.5 Results and Validation

We show here the result for a single node, the wikipedia page entitled *Chess Boxing* [5]. This page exhibits good results which are easily interpretable and can be validated by hand.

For the "Chess Boxing" node, the algorithm iterated over 3000 nodes chosen at random from the nodes between the 100^{th} and the 10.000^{th} best ranked nodes leads to 770 groups of nodes. Figure **??** shows a successful trial leading to the identification of a group along with an unsuccessful trial.

Figures **??** shows the Jaccard similarity matrix of the 770 unfolded communities before cleaning. The columns and lines of the matrix have been rearranged so that columns corresponding to similar groups are next to each other. We see that there are 716 communities very similar to one another, while not similar to

[4] For two sets A and B, the Jaccard similarity is given by $Jac(A,B) = \frac{|A \cap B|}{|A \cup B|}$.

[5] ChessBoxing is a sport mixing Chess and Boxing in alternated rounds.

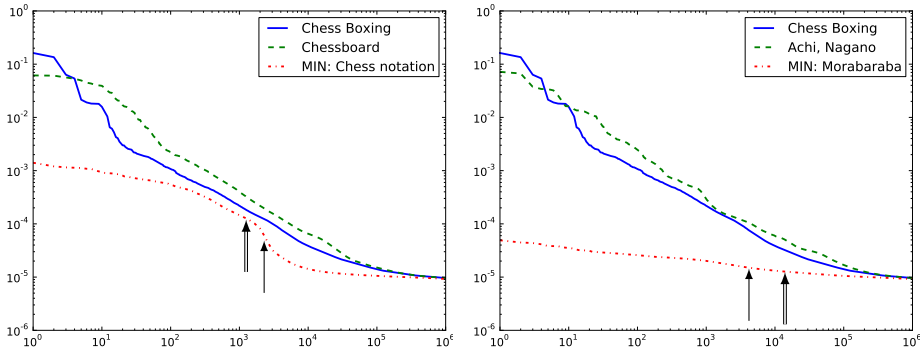


Figure 13: Each figure shows the curves corresponding to a trial: the y axis represents the scores and the x axis represents the ranking of the nodes according to their scores. The first (resp. second) curve is the carryover opinion run from the node Chess Boxing (resp. a candidate for v, the legend shows the label of the candidate), while the third curve shows the minimum, the label of the first ranked node is in the legend. The first trial is successful, while the second is not (no plateau/decrease structure). The double arrow shows the position of the "Chess Boxing" node, while the simple arrow shows the position of the sharpest slope.

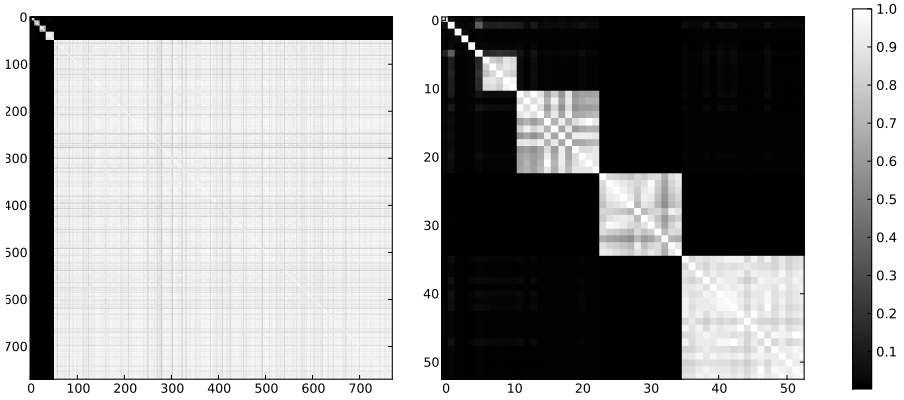
the other ones. If u is in or around a large community, we have a high probability to unfold it, and this probability increases with the size of the community. A problem of the algorithm is that if very large communities exist, the algorithm can have some difficulty to unfold other small communities. We will come back to that problem in the future work section.

When zooming on the rest of the matrix, figure **??**, we see 4 medium size groups of communities and 6 groups containing only a single community: these are actually mistakes of the plateau/decrease detection part of the algorithm and these groups are automatically deleted during the cleaning step.

This decomposition into 5 main groups is easily obtained by intersecting similar groups (we used a Jaccard similarity threshold of 0.7, while the other six singleton groups are automatically deleted. The labels and sizes of the 5 groups are "Enki Bilal" (35 nodes), "Uuno Turhapuro" (26 nodes), "Da Mystery of Chessboxin' " (254 nodes), "Gloria" (55 nodes) and "Queen's Gambit" (1.619 nodes). As we can see the algorithm identifies groups with very different sizes (from 26 nodes to 1.619 nodes on this example) which is a positive feature since other approaches are quite often limited to small sized communities.

Some labels are intriguing, however by checking their meanings on wikipedia on-line, all of them can be justified very easily:

- Enki Bilal is a French cartoonist. Wikipedia indicates that "Bilal wrote [...] Froid Équateur [...] acknowledged by the inventor of chess boxing, Iepe Rubingh as the inspiration for the sport". The nodes in this group



(a) (b)

Figure 14: Figure **??** is the rearranged Jaccard similarity matrix of these 770 communities. We see that there are 716 communities very similar to one another while not similar to the rest of the communities (the big white square). Figure **??** shows a zoom on the top left corner of the matrix.

are mostly composed of its other cartoons.

- Uuno Turhapuro, is a Finnish movie. It is, as Enki Bilal, also acknowledged as the inspiration of the sport, with a scene "where the hero plays blindfold chess against one person using a hands-free telephone headset while boxing another person". The nodes in this group are mostly other cartoons characters or actors in the movies or strongly related to Finnish movies.

- "Da Mystery of Chessboxin' " is a song by an American rap band: "The Wu-Tang Clan". The nodes in the community are related to the band and rap music, which is also relevant.

- "Gloria" is a page of disambiguation linking to many pages containing Gloria in their title. The current wikipedia page of "Chess Boxing" contains the sentence "On April 21, 2006, 400 spectators paid to watch two chess boxing matches in the Gloria Theatre, Cologne". However there is no hyperlink to the page "Gloria Theatre, Cologne" which is a stub. Looking at the records of wikipedia, we found that a link towards the page Gloria was added to the page "Chess Boxing" on May, the 3 2006 and then removed on January, the 31 2008. Due to the central nature of the page "Gloria" within the Gloria community, "Chess Boxing" was part of the Gloria community between these two dates, i.e., when the dataset was compiled!

- Finally, "Queen's Gambit" is a famous Chess opening and the community is composed of Chess related nodes. Even though we could have liked to

label this community "Chess", "Queens' Gambit" is very specific to chess and thus characterizes this community very well.

Surprisingly, the algorithm did not find any community related to boxing. This could be a mistake due to the algorithm itself, however the wikipedia page of "Chess Boxing" explains that most chess boxers come from a chess background and learn boxing afterward. They could thus be important within the community of Chess, but less important within the boxing community. Therefore this could explain that the "Chess Boxing" node lies within the community of Chess, but is at the limit of the boxing community.

5.6 Comparison to another approach

As stated in the related work section, there are other methods to find egocentered communities, all of them based on the optimization of a quality function. We compare here shortly our results to the one of **?** which, we believe, is the most advanced quality function approach since it corrects many drawbacks of previous methods.

Quality function techniques, due to the non-convexity of the optimization problem often lead to small communities, while our approach does not suffer from this drawback. We can indeed check this on the previous example for which the approach of **?** finds only two small communities:

- The first one contains 7 nodes: Comic book, Enki Bilal, Cartoonist, La Foire aux immortels, La Femme Piège, Froid-équateur and Chess boxing. This community is strikingly similar to our community labeled "Enki Bilal" and is very relevant.
- The second one contains 5 nodes: Germany, Netherlands, 1991, International Arctic Science Committee and Chess boxing. This second community is not similar to any of the communities we found and does not seem to be particularly relevant.

6 Conclusion and Perspectives

While studying the global overlapping structure of a real-world network is too complex, studying its community structure as a partition is too restrictive. The local overlapping structure around a node (egocentered community structure) is a good compromise between simplicity and realism. Trying to unfold these egocentered communities by optimising a quality function often leads to poor results, because the optimization landscape is highly non-convex and the optimization often ends up in local minima. In this chapter, we have suggested to look for irregularities in the decrease of a proximity measure in order to avoid these local minima. While our framework is independent of the chosen proximity measure, we still have suggested a new one called the carryover opinion. It has good properties for this application: it is fast to compute, not too simple and parameter-free.

In large graphs, the decrease of the carryover opinion often follows a scale-free law, because a node often belongs to many communities overlapping, fuzzily defined and of different sizes. In that case, no scale can be extracted from the measure and this first approach is limited. Nevertheless this proximity shows how likely it is for two nodes to share at least one community. It also allows to see whether the node characterizes a community by itself (a plateau/decrease structure), is in the middle of a few communities (wavy power-law) or in a middle of many communities (quasi-perfect power-law).

To cope with this limitation we introduced the concept of *multi-egocentered communities*: while a node often belongs to many communities, a well-chosen small set of nodes can characterize a single community. Following this idea, we introduced an algorithm which, given a node, finds all communities centered on that node. Contrary to other existing algorithms, ours avoids local minima, find communities of various sizes and densities, and also allows to label the obtained communities. The algorithm is time efficient and is able to deal with very large graphs. We validated the results on toy graphs, benchmarks and a practical example using a real-world very large graph extracted from wikipedia.

Still, some features of the algorithm can be improved. For instance the detection of irregularities only returns the sharpest decrease. It would be good to find all relevant irregularities, which would provide multi-scale communities.

Furthermore, the algorithm is only looking for bi-centered communities, and some communities might appear only when centered on 3 or more nodes. It would be interesting to incorporate this feature. However it will increase the running time of the algorithm, especially because of unsuccessful trials. More advanced selection of candidates thus needs to be developed. We could for instance add the following criterion: if a candidate is chosen for v, nodes very similar to this candidate might be neglected since they would probably lead to the same result. The speed of the algorithm is a very important feature and is central to make it practical for the study of evolving communities.

As we saw, the algorithm can have some difficulties to find very small communities if there exist very large ones around the node of interest. This might be the reason why when applied on a globally popular node, like "Biology" or "Europe" in the wikipedia network, the algorithm only returns one very big community, while we expected the communities of various sub-fields of Biology or European country related topics. This is a feature of the algorithm that should be improved: relaunching the algorithm again on the induced subgraph of the nodes belonging to the large community detected, or removing the nodes belonging to the big communities from the graph and running the algorithm again should be investigated.

In this book chapter we mainly have focused on a single application of the concept of multi-egocentered communities, which is to unfold all egocentered communities of a node of interest through unfolding its multi-egocentered communities using some other well chosen candidates. At least two other applications of multi-egocentered communities are straightforward and currently under investigation, which are (i) unfolding all nodes of a community given only some

of its members and (ii) unfolding all (overlapping) communities of a network (through unfolding multi-egocentered communities of many small sets of nodes).

This notion of multi-egocentered community could also help to study communities in evolving networks, while the definition of weighted-multi-egocentered communities (potentially with negative weights) should refine the technique even more. These two extensions are currently under investigation.